DOCUMENT RESUME

ED 320 770 SE 051 484

AUTHOR Sudweeks, Richard R.; Tolman, Richard R.

TITLE The Use of Empirical versus Subjective Procedures for Identifying Science Test Items which Function Differentially for Females and Males.

PUB DATE Apr 90

NOTE 30p.; Paper presented at the Annual Meeting of the National Association for Research in Science Teaching (63rd, Atlanta, GA, April 8-11, 1990).

PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.

DESCRIPTORS *Elementary School Science; Elementary Secondary Education; Females; Grade 5; Item Bias; Science Education; *Science Tests; *Secondary School Science; *Sex Bias; *Statistical Analysis; Test Bias; Test Construction; *Test Items

ABSTRACT

The purpose of this study was to compare empirical and subjective procedures for identifying science test items which discriminate against males or females in a fifth grade population (n=926). Techniques such as the Mantel-Haenszel procedure enable test developers to use both a priori and ex post facto procedures to identify specific items that may be functioning with gender bias. The test used in this study was developed as an end-of-year, criterion-referenced instrument and was based on the standards and objectives in the Utah State Science Core Curriculum for the fifth grade. All items were multiple-choice; some used illustrations, but most were text only. The test was initially administered to approximately 300 fifth-grade students. Items were compared that were identified as functioning differentially toward gender. Results reported in this study include: (1) the two methods used led to different results; and (2) several items identified by these methods were more difficult for females. Suggestions for the use of each technique are provided, together with the characteristics of items that discriminated on the basis of gender. (CW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED320770

The Use of Empirical Versus Subjective Procedures for Identifying Science Test Items Which Function Differentially for Females and Males

A paper presented at the
annual meeting of the
National Association for Research in Science Teaching

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Richard R. Tolman
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

by
Richard R Sudweeks
Richard R. Tolman
Brigham Young University

Atlanta, Georgia
April, 1990

SE051484

BEST COPY AVAILABLE

The Use of Empirical Versus Subjective Procedures for Identifying Science Test Items Which Function Differentially for Females and Males

Responsible test developers and users recognize that test scores are influenced by sources of variance other than the construct a test was designed to measure. Because of these extraneous factors, differences in the test scores do not accurately reflect differences in the examinees' true achievement. If this were not the case, test scores would always be completely valid and reliable. The effect of these extraneous factors may vary systematically from one subpopulation of examinees to another. When this occurs, an item is said to function differentially. Such differential item functioning (DIF) can be summarized statistically but is not readily detectable by visual inspection of the items. A visual inspection can detect language or other material that is offensive to one subpopulation or another of the entire population. Eliminating this possibly offensive material helps ensure that different subpopulations are responding to the test items under conditions that are as equal as possible. Thus it is important that test developers conduct both a visual inspection of the items and an empirical study to determine how different subpopulations perform on each specific item.

Methodology

Instruments

The instrument used in this study consisted of 78 multiple-choice items, each including four or five options. The test was designed to assess end-of-year student achievement of the objectives specified in the science component of the Utah Core Curriculum. The instrument was developed by the researchers under a contract with the Utah State Office of Education.

Each item was reviewed by subject matter experts to determine the accuracy of the content and by practicing teachers to ensure content validity. Each item was also reviewed initially by the team of developers for potential gender sensitivity. The items were revised once, based on this initial review. They were later revised on the basis of traditional item analysis statistics summarizing the results of a field test in which the items were administered to more than 300 fifth-grade students selected from four suburban and rural Utah school districts.

Sample

The revised test was administered to more than 1,000 fifth-grade students in 6 suburban and rural school districts in Utah. Administration occurred about three weeks prior to the end of the school year. There were 139 students who were excluded from the present analysis because they failed to indicate their gender on their test form. The results described below reflect the responses

of the remaining 926 fifth grade students. This sample included 460 males (49.7 percent) and 466 females (50.3 percent).

Procedures

The test booklets were examined by gender bias reviewers at the Mountain West Centers for Equity, a federally-sponsored agency. The reviewers critically examined each of the items in the test in search of items containing words, phrases, symbols, or stereotypes which might be offensive to some identifiable subgroup of the population.

In addition to the review by trained, experienced equity specialists, two kinds of statistics were computed for each of the 78 items in the test. The first statistic represented the difference in the difficulty level of each item for the male and female subpopulations. This statistic was calculated by computing the percent of the females who answered an item correctly and the percent of the males who answered correctly. The difference between these two values was then computed by subtracting the males' percentage correct from the females' percentage correct. This difference statistic is commonly referred to in the testing literature as an item's impact (Holland & Thayer, 1988). This statistic is easily computed and has intuitive appeal, but previous researchers have criticized its use as an indicator of item bias (Hunter, 1975; Holland & Thayer, 1988; Dorans, 1989).

Dorans (1989) presents an example which illustrates the problems of using impact as an indicator and emphasizes the need

to investigate the degree to which items function differentially for subpopulations that have been matched in terms of ability level. In recent years, the term differential item functioning (DIF) has become popular as a means of referring to items that are systematically easier (or more difficult) for one subpopulation than for another.

At least three different methods are currently used to study DIF. These include (1) item response theory (Hambleton & Rogers, 1989; Thissen, Steinberg & Wainer, 1988); (2) the Mantel-Haenszel method (Holland & Thayer, 1988); and (3) the standardization method (Dorans, 1989). Scheuneman & Bleistein (1989) provide a helpful description and comparison of these alternative methods.

The Mantel-Haenszel method was used as the method for analyzing the results of this study. This procedure involves classifying the examinees from the two comparison groups (males and females) into matched subgroups based on their total scores on the science test. This matching was accomplished by cutting the distribution of total scores into twelve intervals. Within each interval, the odds of answering the question correctly are computed separately for the males and for the females.

The Mantel-Haenszel procedure is implemented in two steps. The first step involves computation of a chi-square statistic for the 2x2x12 contingency table for each item. The resulting chi-square statistic is used as an initial flagging criterion to identify items which function differentially. Nineteen of the 78 items exceeded the critical value of chi square (alpha=.05) and

were flagged as potential DIF items as a result of this initial step.

Each examinee's total score was then recomputed, based on the remaining 61 items. This purified score was then used as the basis for matching the two comparison groups in the second stage of the analysis.

The statistic produced for each item in the second step of the Mantel-Haenszel procedure is a weighted odds ratio which compares the odds of answering the item correctly for members of the reference group (males in this study) versus the odds of answering the item correctly for members of the focal group (females in this study), averaged across all twelve of the matched score intervals. The value of this weighted average is 1.0 for items which do not function differentially. Items having a value greater than 1.0 for this statistic are differentially easier for males, while items having a value less than 1.0 are differentially easier for females.

Holland and Thayer (1988) proposed a log transformation of the weighted odds ratio which they refer to as delta. Educational Testing Service routinely uses this transformation in their applications of the Mantel-Haenszel procedure. This transformed statistic has a value of zero when there is no DIF in an item. Negative values of delta denote items which are differentially easier for males and positive values denote items that are differentially easier for females. Both the weighted odds ratio

and the delta statistic were utilized to report the results of this investigation.

Results

Initial results indicate that the performance of fifth grade males and females was similar on the 78 item test. Table 1 contains the n, mean, standard deviation, and range for the 460 males and 466 females in the study. The mean score for males was 36.0; the mean score for females was 34.6. The standard deviation of 11.4 for the males was slightly larger than the standard deviation of 9.8 for the females. The range of 6-69 for the male scores was also slightly larger than the range of 7-62 for the female scores. Figure 1 shows the distribution of scores for the males and females.

----- ----- ----- ----- ----- -----

Insert Table 1 About Here

----- ----- ----- ----- ----- -----

----- ----- ----- ----- ----- -----

Insert Figure 1 About Here

----- ----- ----- ----- ----- -----

Table 2 summarizes the results of the item analyses for all 78 items in the test. This table reports the impact for each item as well as the results of the Mantel-Haenszel analysis.

----- ----- ----- ----- ----- -----

Insert Table 2 About Here

----- ----- ----- ----- ----- -----

Fifty of the items have negative impact values, indicating that a larger percentage of males answered the item correctly, while the remaining 28 items are positive, indicating that a larger percentage of females responded correctly. The absolute value of these differences ranges from 0.1 percent to 16.4 percent. Thirty-eight of these differences are less than 3 percent, while 20 others range from 3.01 percent to 5.0 percent, and the remaining 20 exceed 5.0 percent. The column at the far right of Table 2 indicates the degree of DIF in each item, according to the classification procedure used by Educational Testing Service (Dorans, 1989, personal communication). ETS classifies items into three levels of DIF, as shown below.

Category	Amount of DIF	Absolute Value of Delta
A	Nonsignificant	less than 1.0
B	Moderate	between 1.0 and 1.5
C	Heavy	greater than 1.5

Note that nine of the 78 items in the test are classified as manifesting DIF. One of these (#48) is classified as possessing heavy DIF. Inspection of the delta values for these 9 flagged items reveals that 8 of them were differentially easier for the males while one was easier for the females.

All nine of these flagged items have absolute values of impact exceeding 6 percent. However, the remaining 13 items with such relatively high impact are not classified as functioning differentially when male and female subpopulations are matched in terms of ability.

The Mantel-Haenszel procedure is useful for identifying items which function differentially, but it doesn't pinpoint the reasons for the observed DIF. One can only infer possible explanations by examining the individual items. In order to facilitate this examination, the 78 items in the test were classified by the objectives they were intended to assess. The results of this classification are shown in Table 3. Note that five of the nine items possessing moderate or heavy DIF correspond to one of these eight objectives. This is the first category of objectives, which

----- ----- ----- ----- ----- -----

Insert Table 3 About Here

----- ----- ----- ----- ----- -----

is related to animals and ways in which they adapt to the environment. In addition, two of the nine DIF items correspond to

the third category of objectives, which is related to states and forms of energy, plus renewable and non-renewable resources. Two additional categories of objectives each contained only one item with significant DIF. The other four categories of objectives had no items with significant DIF.

The five items corresponding to the first objective category are displayed in Table 4. All five of these items refer to

----- ----- ----- ----- ----- -----

Insert Table 4 About Here

----- ----- ----- ----- ----- -----

characteristics of animals and all five are differentially easier for males. Perhaps this means that objectives related to knowledge of animal characteristics tend to be easier for males than females. However, the test contained seven other items designed to assess this objective category and none of them possessed significant DIF.

Table 5 shows the remaining three items that were differentially easier for males. Two of these items relate to forms of energy and the third relates to the distinction between the terms "revolution" and "orbit." Item 17, shown in the middle of Table 5, is the item having the largest absolute value of impact.

----- ----- ----- ----- ----- -----

Insert Table 5 About Here

----- ----- ----- ----- ----- -----

The one item showing significant DIF in favor of females is shown in Table 6. This was the most difficult item in the entire test. Only nine percent of the males answered this item correctly while 15 percent of the females gave a correct response. Examination of the response patterns of the two

---

Insert Table 6 About Here

---

subpopulations reveals that three of the four distractors were more attractive for members of both groups than the correct answer. We suspect that this item was simply too difficult for most of the students and they simply responded in a random manner. Note that this item focuses on two ideas--the effect of a convex lens on the size of the image that is formed and the effect of a convex lens on the light rays. We hypothesize that this item would better assess the knowledge of fifth-graders if it focused on only one of these two ideas.

Some of the items in the test had positive impact values exceeding 6 percent, meaning that 6 percent more females than males responded correctly. However, when the groups were matched in terms of science ability, this apparent difference disappeared and the items did not show significant DIF. Two examples of such items are shown in Table 7. We initially suspected that item 10 would be differentially easier for female fifth-graders because it refers

to knowledge associated with reproduction. We hypothesized that fifth grade females would be more knowledgeable about this subject than fifth grade males because of differences in their physical development and maturation. If we looked only at the impact statistic, this hypothesis would appear to be supported. However, when the groups are matched in ability, the item no longer shows significant DIF.

----- ----- ----- ----- ----- -----

Insert Table 7 About Here

----- ----- ----- ----- ----- -----

The amount of DIF in an item is a matter of degree. However, some of the items in the test displayed almost complete absence of DIF. Table 8 presents three of the items which illustrate this phenomenon. Notice that both the delta and the impact statistics for each item are very close to zero and that the value of the weighted odds ratio (alpha) is very close to 1.0.

----- ----- ----- ----- ----- -----

Insert Table 8 About Here

----- ----- ----- ----- ----- -----

Two additional items with relatively large values of impact favoring the males are displayed in Table 9. Both of these items refer to objectives related to knowledge of the principles of magnetism. We suspected that these items would show DIF in favor

of the males, but when the groups were matched in terms of ability, neither of these two items revealed significant DIF. In other words, when the male and female populations are matched in terms of science ability, the two groups perform equally well.

----- ----- ----- ----- ----- -----

Insert Table 9 About Here

----- ----- ----- ----- ----- -----

Item 25, shown in Table 9, is the only one of the 78 items in the fifth grade test which was flagged by the sensitivity reviewers as being potentially biased. They objected to inclusion of the term "curling iron" as distractor E and suggested that this term be replaced with a more generic term. This concern is similar to concerns raised in two other tests at different grade levels we had constructed. In the second grade test the reviewers suggested that the word "snowman" be replaced with the term "snow sculpture." In another test, the sensitivity reviewers expressed concern about use of the term "bullfrog," when used in reference to a particular species of frog. We have not yet computed DIF statistics for these two items from the other grade levels. However, the inclusion of the term "curling iron" in item 17 did not cause this item to function differentially. The frequency distributions of responses of the males and females to this item, as shown in Table 9, appear to support this interpretation. Distractor E was less attractive to both groups than the correct answer and two of the other

distractors. Thus the phrase "curling iron" did not function as a strong distractor for members of either group.

Discussion and Conclusions

The two methods used in this study to detect biased items led to different results. The single item identified by the sensitivity review was not flagged by the empirical procedure. Conversely, the nine items identified by the Mantel-Haenszel procedure were not flagged by the three sensitivity reviewers. This finding is not surprising since the two methods address different issues and serve different purposes. We believe that responsible test developers should utilize both procedures. Pre-publication versions of tests should be subjected to the sensitivity review process prior to field testing. However, it is still important to determine empirically the degree to which individual items function differentially within the two subpopulations. Statistical analysis of the distribution of response patterns of the two groups is necessary to estimate DIF.

Our results also confirm the need to match the comparison groups in terms of ability levels before performing statistical analyses. The impact statistic indicated that many items functioned differentially which were not flagged by the Mantel-Haenszel procedure. We believe that stopping the analysis after calculating the impact of each item is premature.

The results of this study raise several important questions which are unresolved in our minds. Eight of the nine flagged items

were more difficult for females. Why? What factors caused these items to function differentially for males and females? Several factors may account for these observed differences. These possible factors include, but are not limited to (a) differences in the students' reading experiences and interests; (b) differences in their academic experience; and (c) differences in the out-of-school experiences of the two groups.

A second issue which needs to be addressed focuses on the question of what should be done with the flagged items. Should they be eliminated summarily? The one item possessing heavy DIF should probably be deleted. What about the other four flagged items measuring this same objective category? If all or most of the items measuring a given objective show evidence of DIF, should that objective be deleted from a test? Does this mean that the objective functions differentially? This issue is a curricular question which is not easily answered.

Since there were twelve items intended to measure the first objective category, the four items showing evidence of moderate DIF could probably be eliminated without impairing the content validity of the test. But, what if these were the only items measuring that objective? Should items which show evidence of DIF be eliminated, if doing so reduces content validity?

Should items which contain familiar terms like "snowman" be revised to incorporate less offensive terms that are less familiar to the target group of examinees for whom the test is designed? Should "snowman" be replaced with "snow sculpture" if there is good

reason to assume that the substitute term is not likely to be in the vocabulary of the intended examinees?

Table 1. Science Achievement Scores for Male and Female Fifth Graders

Group	*n*	Mean	s.d.	Range
Males	460	36.0	11.4	6-69
Females	466	34.6	9.8	7-62
	926			

Table 2. Item Statistics

Item	Objective	Difficulty Level			Chi Square	Common Odds Ratio	Delta	DIF Category
		Females	Males	Impact				
3	1-1	74.2	82.5	-8.30	7.62	1.63	-1.14	B
4	1-1	50.3	52.0	-1.70	0.01	0.99	0.02	A
5	1-2	76.8	84.1	-7.30	8.22	1.65	-1.17	B
6	1-3	44.5	60.4	-15.90	16.64	1.84	-1.44	B
7	1-4	40.9	54.2	-13.30	10.70	1.68	-1.22	B
8	1-5	19.6	20.3	-0.70	0.15	0.94	0.14	A
9	2-1	85.6	83.2	2.40	1.07	0.86	0.35	A
10	2-2	55.5	47.7	7.80	6.33	0.70	0.83	A
11	2-3	64.4	58.3	6.10	6.48	0.70	0.83	A
12	2-4	42.5	42.1	0.40	1.07	0.89	0.28	A
13	2-5	24.9	31.3	-6.40	4.30	1.33	-0.66	A
14	2-6	61.9	57.4	4.50	3.92	0.75	0.67	A
15	3-1	19.8	22.3	-2.50	0.17	1.11	-0.24	A
16	3-2	23.1	25.8	-2.70	0.00	1.06	-0.13	A
17	3-3	35.7	52.1	-16.40	18.57	1.88	-1.48	B
18	3-4	27.3	26.6	0.70	0.72	0.93	0.16	A
19	3-5	38.4	40.9	-2.50	0.07	1.07	-0.16	A
20	4-2	48.3	58.8	-10.50	5.86	1.51	-0.97	A
21	4-6	68.2	65.9	2.30	2.20	0.85	0.38	A
22	5-1	43.6	45.5	-1.90	0.02	1.02	-0.05	A
23	5-1	24.2	30.6	-6.40	2.50	1.32	-0.65	A
24	4-3	20.4	22.1	-1.70	0.43	1.10	-0.23	A
25	4-5	31.2	37.7	-6.50	2.00	1.28	-0.58	A
26	5-2	12.4	17.0	-4.60	2.42	1.39	-0.77	A
27	5-3	53.1	52.5	0.60	0.28	0.94	0.14	A
28	5-5	42.0	43.4	-1.40	0.05	0.97	0.07	A
29	6-1	35.0	30.8	4.20	3.64	0.76	0.63	A
30	5-2	38.9	40.3	-1.40	0.00	1.04	-0.08	A
31	6-2	78.4	77.3	1.10	0.19	0.97	0.08	A
32	6-3	15.3	8.6	6.70	8.78	0.52	1.51	B
33	6-4	47.6	40.2	7.40	5.82	0.72	0.79	A
34	6-5	32.5	36.2	-3.70	0.07	1.14	-0.30	A
35	7-1	41.3	53.6	-12.30	8.97	1.64	-1.17	B
36	7-2	38.6	43.4	-4.80	0.38	1.15	-0.33	A
37	7-4	50.5	51.2	-0.70	0.20	0.93	0.18	A
38	7-4	36.4	40.7	-4.30	0.15	1.08	-0.18	A
39	8-3	50.0	49.5	0.50	0.62	0.91	0.23	A
40	7-5	24.9	29.4	-4.50	1.05	1.18	-0.40	A
41	5-4	33.8	35.5	-1.70	0.01	1.02	-0.05	A
42	7-6	54.2	54.3	-0.10	0.43	0.96	0.11	A

Table 2. Item Statistics (continued)

Item	Objective	Difficulty Level			Chi Square	Common Odds Ratio	Delta	DIF Category
		Females	Males	Impact				
43	8-1	61.0	62.5	-1.50	0.01	1.01	-0.02	A
44	8-3	67.9	68.3	-0.40	0.00	1.03	-0.06	A
45	8-3	70.7	74.1	-3.40	1.37	1.33	-0.67	A
46	1-1	84.0	84.5	-0.50	0.12	1.14	-0.23	A
47	1-1	50.1	47.9	2.20	1.70	0.85	0.38	A
48	1-2	71.4	82.9	-11.50	19.13	2.27	-1.92	C
49	1-3	53.9	50.7	3.20	0.91	0.87	0.32	A
50	1-4	68.2	65.9	2.30	1.99	0.86	0.36	A
51	1-5	60.9	59.0	1.90	1.43	0.86	0.34	A
52	2-1	32.8	34.9	-2.10	0.05	1.00	0.00	A
53	2-2	32.1	36.1	-4.00	0.18	1.11	-0.25	A
54	2-3	50.7	47.5	3.20	1.78	0.85	0.38	A
55	2-4	34.6	38.9	-4.30	0.47	1.18	-0.38	A
56	2-5	45.0	45.5	-0.50	0.32	0.98	0.04	A
57	2-6	70.0	66.9	3.10	3.18	0.77	0.62	A
58	3-1	24.6	21.8	2.80	0.99	0.86	0.36	A
59	3-2	31.8	44.5	-12.70	8.86	1.62	-1.13	B
60	3-3	38.0	44.9	-6.90	3.68	1.29	-0.61	A
61	3-4	36.9	28.7	8.20	7.11	0.69	0.85	A
62	3-5	37.7	34.9	2.80	0.84	0.88	0.30	A
63	4-2	58.0	57.8	0.20	0.34	0.91	0.22	A
64	4-3	42.0	48.0	-6.00	0.30	1.19	-0.41	A
65	4-5	32.8	37.5	-4.70	1.06	1.22	-0.46	A
66	4-6	33.4	43.0	-9.60	6.46	1.49	-0.93	A
67	5-2	28.3	28.0	0.30	0.48	0.91	0.23	A
68	5-5	36.7	40.3	-3.60	0.00	1.04	-0.08	A
69	6-1	45.3	47.7	-2.40	0.07	1.01	-0.02	A
70	6-2	43.0	43.9	-0.90	0.65	0.93	0.17	A
71	6-3	18.7	20.0	-1.30	0.01	1.06	-0.14	A
72	6-4	47.3	43.4	3.90	3.97	0.77	0.62	A
73	6-4	44.2	40.3	3.90	2.73	0.80	0.52	A
74	7-1	25.3	29.8	-4.50	0.22	1.11	-0.24	A
75	7-3	42.6	48.2	-5.60	0.77	1.18	-0.38	A
76	7-5	36.6	38.0	-1.40	0.00	0.97	0.07	A
77	7-6	40.3	43.2	-2.90	0.00	1.03	-0.08	A
78	8-1	52.8	56.3	-3.50	0.20	1.12	-0.27	A
79	8-2	61.1	55.0	6.10	4.92	0.76	0.65	A
80	8-2	74.1	72.4	1.70	0.89	0.89	0.26	A

Table 3. Number of Items Which Function Differentially Classified by Direction and Degree of DIF

Category	Objectives/Concepts	Total Number of Items	Direction	Degree of DIF: Not Significant	Moderate	High
1	Characteristics of animals & their adaptations to the environment	12	Favors Males	1	4	1
			Favors Females	6	0	0
2	Characteristics of plant parts; plant reproduction	12	Favors Males	4	0	0
			Favors Females	8	0	0
3	States and forms of energy; renewable & nonrenewable	10	Favors Males	4	2	0
			Favors Females	4	0	0
4	Magnetism and electricity	8	Favors Males	6	0	0
			Favors Females	2	0	0
5	States of matter; atomic and molecular characteristics	9	Favors Males	6	0	0
			Favors Females	3	0	0
6	Characteristics of light; mirrors and lenses	10	Favors Males	3	0	0
			Favors Females	6	1	0
7	Movements of the Sun, Earth and Moon; Causes of night, day and seasons	10	Favors Males	6	1	0
			Favors Females	3	0	0
8	Conservation of natural resources; pollution	7	Favors Males	4	0	0
			Favors Females	3	0	0
		78	Total Favoring Males	34	7	1
			Total Favoring Females	35	1	0
				69	8	1

Table 4. Items With Significant DIF Which Are Easier for Males

Item 48.

M%	F%		Which of these animals hunts in groups?	
5	9	A.	hawk	Impact = -11.5%
83	71	*B.	wolf	Alpha = 2.27
6	9	C.	rabbit	Delta = -1.92
6	10	D.	cougar	DIF = C

Item 6.

M%	F%		Which of these animals migrates?	
60	45	*A.	duck	Impact = -15.9%
24	33	B.	bear	Alpha = 1.85
8	11	C.	snake	Delta = -1.44
7	11	D.	skunk	DIF = B

Item 5.

M%	F%		Which of these animals lives in colonies?	
9	11	A.	salmon	Impact = -7.3%
3	6	B.	shark	Alpha = 1.65
84	77	*C.	ant	Delta = -1.17
3	5	D.	octopus	DIF = B

Item 7.

M%	F%		Which of these animals became extinct because of overhunting?	
54	41	*A.	passenger pigeon	
9	8	B.	quail	Impact = -13.3%
27	36	C.	buffalo	Alpha = 1.68
10	15	D.	panda bear	Delta = -1.22
				DIF = B

Item 3.

M%	F%		Which of these has a break-away tail to help it escape from predators?	
3	2	A.	zebra	
10	16	B.	snake	Impact = -8.3%
4	8	C.	sparrow	Alpha = 1.63
83	74	*D.	lizard	Delta = -1.15
				DIF = B

Table 5. Additional Items With Significant DIF Which Are Easier for Males

Item 59.

M%	F%		Which of these refers to fossil energy?
25	28	A.	energy obtained from hot underground water or steam
45	32	* B.	energy obtained by burning oil, natural gas and coal
12	15	C.	electrical energy generated by turbines turned by water
10	16	D.	light energy converted to electricity by a solar battery
8	9	E.	energy obtained from burning garbage

Impact = -12.7%
Alpha = 1.62
Delta = -1.13
DIF = B

Item 17.

M%	F%		Which energy form is produced by a generator?
15	15	A.	light
18	34	B.	heat
11	11	C.	mechanical
52	36	*D.	electrical
4	5	E.	chemical

Impact = -16.4%
Alpha = 1.88
Delta = -1.48
DIF = B

Item 35.

M%	F%		Which of these describes the word orbit as it refers to the movement of objects in space?
38	33	* A.	the path followed by one object traveling around another object in space
24	23	B.	the movement of an object away from Earth
20	22	C.	the spinning movement of an object around an imaginary center like a wheel
9	7	D.	the path used by astronauts to go to the moon

Impact = -12.3%
Alpha = 1.64
Delta = -1.17
DIF = B

Table 6. Items With Significant DIF Which Favor Females

Item 32.

M%	F%		What happens to light when it passes through a convex lens?
38	33	A.	parallel light rays spread out
24	33	B.	the images it forms appear larger
20	22	C.	a spectrum appears on the other side of the lens
9	7	D.	the images it forms practically disappear
9	15	* E.	the images it forms appear smaller

Impact = +6.7%
Alpha = 0.52
Delta = +1.51
DIF = B

Table 7. Items Which Appear to be Easier for Girls, But the Amount of DIF Is Nonsignificant

Item 11.

M%	F%		Which of these is a fruit?
58	64	* A.	tomato
7	3	B.	carrot
4	4	C.	potato
8	8	D.	cabbage
22	20	E.	beet

Impact = +6.1%
Alpha = 0.70
Delta = +0.83
DIF = A

Item 10.

M%	F%		What word or phrase is used to describe the uniting of a plant sperm cell with a plant egg cell?
11	10	A.	transformation
10	10	B.	cross-ventilation
21	14	C.	cross-pollination
48	55	* D.	fertilization
11	12	E.	self-pollination

Impact = +7.8%
Alpha = 0.70
Delta = +0.83
DIF = A

Table 8. Items Displaying Almost Complete Absence of DIF

Item 56.

M%	F%	How have some plants adapted to living where they could be eaten by animals?
12	14	A. by catching and digesting insects
46	45	* B. by having spines instead of broad leaves
26	20	C. by having hard and glossy leaves
15	20	D. by having columns of air in their stems

Impact = -0.50%
Alpha = 0.98
Delta = +0.04
DIF = A

Item 43.

M%	F%	Which of these is a renewable resource?
13	13	A. fossil fuel
63	61	* B. trees
10	12	C. oil
14	14	D. natural gas

Impact = -1.5%
Alpha = 1.01
Delta = -0.02
DIF = A

Item 69.

M%	F%	Which of these is transparent?
15	15	A. frosted glass
14	14	B. typing paper
10	9	C. a wall
48	45	* D. glass
14	17	E. wood

Impact = -2.4%
Alpha = 1.01
Delta = -0.02
DIF = A

Table 9. Items in Which Apparent DIF Largely Disappears When the Groups are Matched in Terms of Science Ability

Item 25.

M%	F%	In what tools and machines are electromagnets used?
25	26	A. compasses
38	31	* B. generators
19	20	C. calculators
7	7	D. wrenches
12	16	E. curling irons

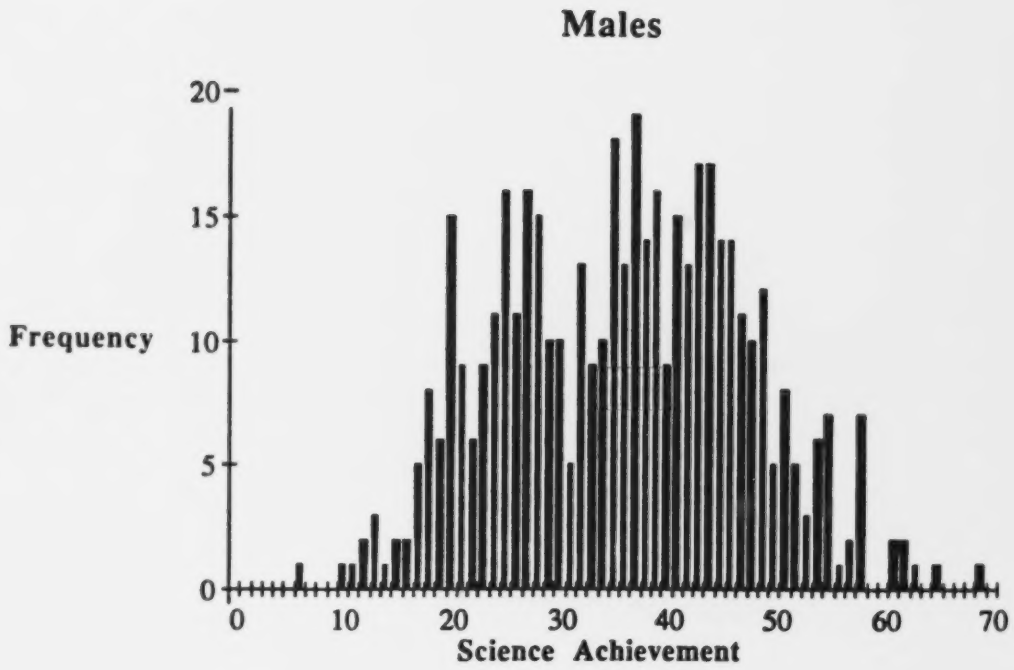
Impact = -6.50%
Alpha = 1.28
Delta = -0.58
DIF = A

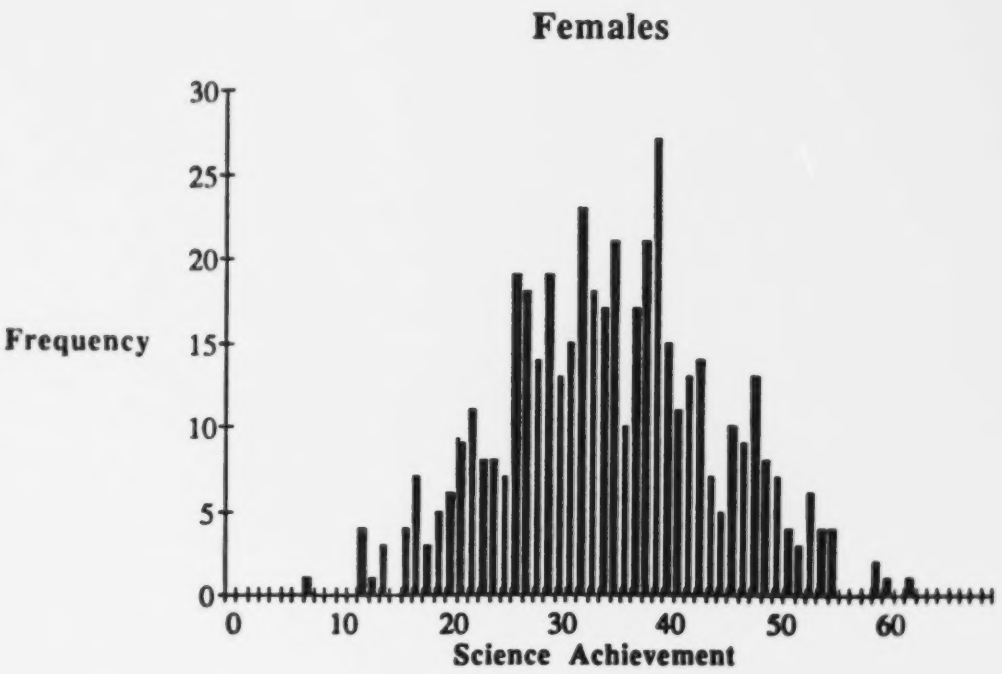
Item 64.

M%	F%	Which of these statements describes electromagnets, but not permanent magnets?
17	15	A. they can be made by moving other permanent magnets across certain metals
48	42	* B. they are made by electric current passing through a coiled wire
19	23	C. they attract certain metals
10	13	D. they become magnetized by being rubbed with a silk cloth
6	7	E. they will not harm computer disks

Impact = -6.0%
Alpha = 1.19
Delta = -0.41
DIF = A

Figure 1. Distribution of Science Test Scores for Fifth Grade Males and Females.





References

Berk, R.A. (Ed.) (1982). Handbook of methods for detecting test bias. Baltimore: Johns Hopkins University Press.

Crocker, L. & Algina, J. (1986). Introduction to classical and modern test theory. New York: Holt, Rinehart & Winston.

Dorans, N.J. (1989). Two new approaches to assessing differential item functioning: Standardization and the Mantel-Haenszel method. Applied Measurement in Education, 2, 217-233.

Educational Testing Service, (1987). ETS Sensitivity Review Process: An overview. Princeton, NJ: Educational Testing Service.

Green, B.F., Crone, C.R., & Folk, V.G. (1989). A method for studying differential distractor functioning. Journal of Educational Measurement, 26, 147-160.

Hambleton, R.K. & Rogers, H.J. (1989). Detecting potentially biased test items: Comparison of IRT Area and Mantel-Haenszel Methods. Applied Measurement in Education, 2, 313-334.

Holland, P.W. & Thayer, D.T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H.Wainer and H.I. Braun (Eds.), Test Validity (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hunter, J.E.(1975, December). A critical analysis of the use of item means and item-test correlations to determine the presence or absence of content bias in achievement test items. Paper presented at the National Institute of Education Conference on Test Bias, Annapolis, MD.

Kahle, J.B. & Lakes, M.K. (1983). The myth of equality in science classrooms. Journal of Research in Science Teaching, 20, 131-140.

Linn, M.C., De Benedictis, T., Delucchi, K., Harris, A., & Stage, E. (1987). Gender differences in national assessment of educational progress science items: What does "I don't know" really mean? Journal of Research in Science Teaching, 24, 267-278.

Mantel, N. & Haenszel, W.M. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the National Cancer Institute, 22, 719-748.

Scheuneman, J.D. & Bleistein, C.A. (1989). A consumer's guide to statistics for identifying differential item functioning. Applied Measurement in Education, 2, 255-275.